Sigma J Eng & Nat Sci 8 (2), 2017, 155-160

Publications Prepared for the Innovations on Intelligents Systems and Applications Symposium ASYU 2016
Akıllı Sistemler ve Uygulamalardaki Yenilikler Sempozyumu ASYU 2016 için Hazırlanan Yayınlar

Research Article / Araştırma Makalesi

ASSOCIATION RULE FOR CLASSIFICATION OF BREAST CANCER PATIENTS

Tuba PALA*[1], İbrahim YÜCEDAĞ[2], Hasan BİBEROĞLU

[1]*Institute of Science and Technology, Duzce University,DUZCE*
[2]*Computer Engineering Department, Duzce University, DUZCE*

Received/Geliş: 05.12.2016 Accepted/Kabul: 11.12.2016

ABSTRACT

Data mining studies carried out on medical databases are very important in order to make an effective medical diagnosis. The purpose of data mining is to extract information from databases, to define clear and understandable patterns. In this study, an approach was presented to generate association rules on the data of breast cancer patients. Apriori algorithm is used for the extract of the rules. Apriori algorithm is usually used for the market - basket analysis. Apriori algorithm is used to determine customer shopping profiles or to campaign, in order to catch the shopping patterns. In this study, apriori algorithm was used in the extraction of the rules within the medical data. UC-Irvine archive repository of machine learning datasets [1] - Breast Cancer dataset has been studied. This dataset, including 9 attribute and 1 class atribute. It consists of records of 286 patients with 10 attributes. The study was carried out by using the Weka data mining program.
Keywords: Data mining, classification, association rule, breast cancer dataset, medical diagnosis.

1. INTRODUCTION

Breast cancer is the most prevalent type of cancer in humans after lung cancer throughout the world. In the most of both developed and developing countries, it is the most common type of cancer seen in women. It is also the leading cause of death from cancer for women [2]. There has been an increase in the frequency of breast cancer cases since 1970s and modern, western lifestyle is viewed as a factor in this increase. As seen in other diseases and cancer types, early diagnosis is vital in breast cancer cases. If breast cancer is diagnosed in early phases before spreading, the patient can have a high chance for survival at a rate of %96 [3]. Therefore, an analysis using Apriori algorithm from association rules algorithms that present recurring patterns on data from breast cancer patients has been made in this study. Thus, it is aimed to assist those working in the field through the presented rules. Developments in computer technology have led to an increased volume of data and increasing efficiency with this making it clear for a need to analyze data. In recent years, tools, methods and techniques to analyze great volume of data have emerged. Data mining is the most used one of these methods and techniques. As is evident from its name, data

* Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: tubapala@duzce.edu.tr, tel: (545) 244 88 98

mining is a technique that tries to extract valuable and distinct information from the mass data. The goal of data mining is to analyze data gathered in data bases with mathematical and statistical methods and to find out rules, relations, structures or previously unknown information [4]. Data mining is used in healthcare field in many ways. It can make a significant contribution to the healthcare institutions and professionals in many areas such as the diagnosis of the disease, determining treatment, reducing treatment period and making decisions regarding hospital administration [5]. Moreover, it can assist physicians to make decisions based on deductions from previous data or experience. Using data mining in the field of healthcare will be a great help for physicians to make the most appropriate decisions. There are various studies carried out on classification and estimation of patterns in data sets of breast cancer and other medical cases in recent years. Polat and Gunes [6], carried out breast cancer analysis by using LS-SVM (least squares – support vector machine) classification algorithm. The strength of LS-SVM was examined by using accuracy, sensitivity, specificity, K fold cross specification and confusion matrix analyses.The classification accuracy of the method was found %98.53 and when compared to other classification methods, this rate was viewed highly prospective. The results show that the method is effective and a new, smart and assistive system for diagnosis. This study [7] offers an automatic diagnosis system for breast cancer detection based on association rules and artificial neutral networks. Association rules were used in order to reduce the size of data set. Following the reduction of data set quality, comparison of the performances of the artificial neutral networks was carried out by means of the model developed applying artificial neutral networks. In performance evaluation, 3 fold cross validation method was used and it showed that the model developed is a successful model with a validation rate of %95.6.

Akay [8] suggested a SVM (Support Validation Machines) based model and the validation rate of the model was found as %99.51. The significance of each attribute in the breast cancer data set was measured through F-score calculation. Delen at al. [9] developed a predicator model that predicts the survival chances of a breast cancer case by using artificial neural networks, logistic regression and decision tree analysis. They used SEER breast cancer data. The classification validity of three algorithms was compared by using 10 fold cross validation for this data set composed of 433272 patient data and 72 attributes. While decision tree algorithm got the best result with %93.5 accuracy, logistic regression got the lowest value of %89.2. In this study [10], attribute selection phase was formed by combining PCA (principal component analysis) used in finding the relationships between attributes in vast data bases and Apriori algorithm. After this phase, a system was developed by applying NN (artificial neural network). This result is highly prospective compared to accuracy values in other studies. Lavaya at al. [11] carried out a study on the effects of attribute selection analysis on the success of classification accuracy. Jacob at al. [12] carried out a study based on the comparison of the success of the attribute selection on various algorithms. Data set is breast tissue data. It has been concluded that attribute selection is influential in certain algorithms but in some others it is not the case. In this study [13], two rule-based models have been studied. While the first model is a system based on decision trees and fuzzy rules, the second is based on association rules and fuzzy logic rule mining. Both of the models were tried on data from breast cancer patients. With the study, two systems that provide accurate fuzzy rule mining have been offered. It has been concluded that both systems get similar results and input attribute value plays an important role in the classification performance of the results. In this study [14] breast cancer and adult data set have been evaluated according to three calculations which are apriori, entropy and variance. The fact that variance value in breast cancer data is high and entropy value is low shows that rule variety can be large. This study presents that various rules could be created on different data sets adding the entropy and variance parameters along with support and confidence values in association rule practices. Jabbar at al [15] proposed a hybrid model by use of association rules, genetic algorithm, gini-index and statistics values. The results demonstrate that assistive classificatory rules are created in disorder estimation. This study testing various classification algorithms on medical data sets shows that the usage of rules in

diagnosis can be effective and beneficial. Later on, Apriori algorithm has been applied to these data sets. The best, the strongest in other words, association rules found as a result of the algorithm application are given and evaluated in the conclusion and review of the study. The remaining part of the study is organized as follows; In part 2, data set used in the study is introduced and association rules and Apriori algorithm is explained and in the third and last parts the results are presented and the study is evaluated. Moreover, future research is included.

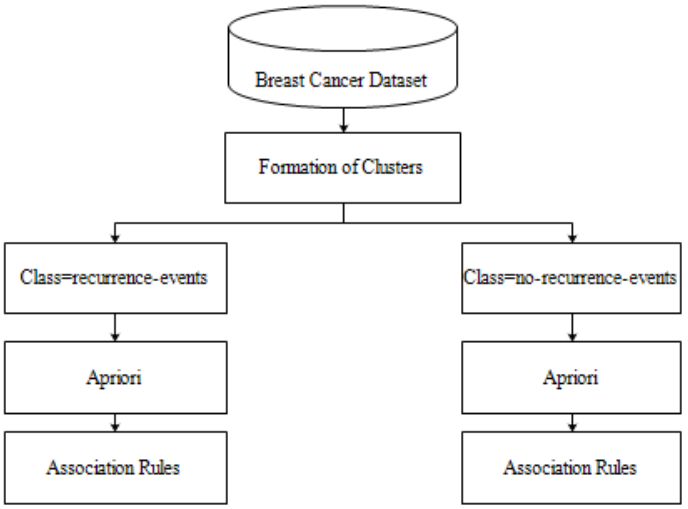

Figure 1. Flowchart of the Proposed Approach

2. METHOD

The data set used in this study was prepared by Institute of Oncology University Medical Center in 1998 and includes information about breast cancer cases [1]. Table 1: Summarized Data Set

Table 1. Breast Cancer Dataset Description.

Dataset	No. of Attributes	No. of Instances
Breast Cancer	10	286

Table 2. Breast Cancer Dataset Attribute Description.

No	Relabeled values	Attribute Description
1	Class	1 – no-recurrence-events 2 – recurrence-events
2	Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3	Menopause	lt40, ge40, premeno
4	tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
5	inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
6	node-caps	yes, no
7	deg-malig	1, 2, 3
8	breast	left, right.
9	breast-quad	left-up, left-low, right-up, right-low, central
10	irradiat	yes, no

The dataset with detailed provision in Table 2 comprise of 286 patients and 10 attributes. The meanings of these data and value ranges are shown in Table 2. In Table 3, the distribution of the class attributes is given.

Table 3. The distribution of class attribute

Class	Frequency	Percent (%)
1 - no-recurrence-events	201	70.2797
2 - recurrence-events	85	29.7203

2.1. Association Rule Algorithm

Association rule is a data mining approach that ensures the discovery of association behavior in data by analyzing data at hand. The discovered patterns show the relationship between the attribute values that mostly co-occur in the data set[17]. In data mining, association rules learning algorithm is highly popular and a well-studied method to discover curious relationships between attributes in vast data bases [18]. Piatetsky – Shapiro defines it as a method presenting strong rules in data bases by using various calculations and analyzing them [19]. In presenting strong rules, the association rule has been used for the discovery of the relationships between products in vast transaction records in Supermarket systems. For example, Rule {onion, potato} => {Burger}, according to a rule extracted from a supermarket transaction record, if a customer buys onions and potatoes then this customer will most probably buy a burger, which is an interpretation of the rule. These rules can be used in decisions regarding marketing activities such as promotion, pricing or product placement. Along with the marketing example above, association rules are utilized in such application areas as Web mining, threat detection and bioinformatics [19]. In order to find out which one of the associations is more valid, a performance analysis is conducted. Two measurements are basically used as performance criteria: support and confidence. Between X and Y ($X \Longrightarrow Y$ for association rule); Support is defined as the probability of co-occurrence of both the initial and result variables. Probability value is shown as the rate of the number of all calculations involving both X and Y to the number of all calculations (1).

$$\text{Support} = \frac{\text{The number of calculations with both X ve Y}}{\text{The number of all calculations}} \quad (1)$$

Confidence is a criterion for the accuracy of the rule and the probability of the result attribute occurring after the initial attribute. Confidence value is a conditional probability criterion and is shown as the rate of the number of calculations involving both X and Y to the number of calculations involving only X (2).

$$\text{Confidence} = \frac{\text{The number of calculations with both X ve Y}}{\text{The number of all calculations}} \quad (2)$$

2.2. Apriori Algorithm

Algorithms basically have a repetitive structure. The aim of algorithms is to find the strongest rules in the database, in other words the most repetitive item sets. To this end, the database is scanned several times. In the first phase, the number of repetitions of the items in the database is defined and this number represents the support value. Items below the support value, that is they are lower than support value, are not included in the item sets. In each scan, previously found common item sets are used to create potential common item sets called candidate item sets. The support value of the candidate item sets are calculated by means of scanning and those with minimum support criterion form the common item sets. Common items become candidate item sets for the next scan. This calculation proceeds until there is no common item sets to be found. The main approach of this algorithm is, if K-item set has the minimum support criterion, subsets of this item set also have the minimum support criterion [17].

3. RESULTS AND EVALUATION

In this study, separate rules were generated on the classes by the Apriori algorithm, which is based on the association rules learning algorithm using the breast cancer data set. The rules are the strongest, or the most frequent, recurrent rules of breast cancer. Table 4 lists the 10 strongest rules for no recurrence in breast cancer patients.

Table 4. Best Rules for Patients under No Recurrence Class

No	Rules	Coverage	Confidence
1	node-caps=no ==> Class=no-recurrence-events	171	100%
2	inv-nodes=0-2 ==> Class=no-recurrence-events	167	100%
3	irradiat=no ==> Class=no-recurrence-events	164	100%
4	inv-nodes=0-2 and node-caps=no ==> Class=no-recurrence-events	160	100%
5	node-caps=no and irradiat=no ==> Class=no-recurrence-events	151	100%
6	inv-nodes=0-2 ==> node-caps=no	167	96%
7	inv-nodes=0-2 and Class=no-recurrence-events ==> node-caps=no	167	96%
8	inv-nodes=0-2 ==> node-caps=no and Class=no-recurrence-events	167	96%
9	node-caps=no ==> inv-nodes=0-2	171	94%
10	node-caps=no and Class=no-recurrence-events==> inv-nodes=0-2	171	94%

When we evaluate some of the rules, the results are as follows. Example, 1. Rule, If there is no lymph node capsule, the disease will not recur 100%. Example, 4. Rule, If the spreading lymph node is in the 0-2 scale and there is no lymph node capsule, the disease does not recur 100%.

Table 5. Best Rules for Patients under Recurrence Class

No	Rules	Coverage	Confidence
1	irradiat=no ==> Class=recurrence-events	54	100%
2	node-caps=no ==> Class=recurrence-events	51	100%
3	breast=left ==> Class=recurrence-events	49	100%
4	menopause=premeno ==> Class=recurrence-events	48	100%
5	inv-nodes=0-2 ==> Class=recurrence-events	46	100%
6	deg-malig=3 ==> Class=recurrence-events	45	100%
7	inv-nodes=0-2 node-caps=no ==> Class=recurrence-events	41	100%
8	node-caps=no irradiat=no ==> Class=recurrence-events	37	100%
9	breast=right ==> Class=recurrence-events	36	100%
10	inv-nodes=0-2 irradiat=no ==> Class=recurrence-events	36	100%

In Table 5, there are 10 strongest rules for recurrence class in breast cancer patients. When we interpret some rules from the rule table, which is recurrence class, the results are as follows. Example, 1. rule, If the patient has not received radiotherapy, the disease will recur 100%. Example, 6. rule, If the malignancy grade of the tumor is 3, the disease will recur 100%. It is thought that this study will be of great help to the physicians and practitioners in interpreting the diseases. The approach of this study can also be used to derive rules for other diseases. In addition, the Predictive Apriori Algorithm (Estimator Apriori) and the Frequent Pattern FP-Growth Algorithm, which are from the association rule algorithms, can be utilized and comparisons can be made by producing rules.

REFERENCES / KAYNAKLAR

[1] http://archive.ics.uci.edu/ml/datasets [last access: 30.05.2016].

[2] Özmen, Vahit, et al. "Türkiye'de Meme Kanseri Erken Tanı Ve Tarama Programlarının Hazırlanması "Sağlık Bakanlığı meme kanseri erken tanı ve tarama alt kurulu raporu"." Meme Sagligi Dergisi/Journal of Breast Health 5.3 (2009).

[3] https://tr.wikipedia.org/wiki/Meme_kanseri (Son erişim: 12.06.2016)

[4] Han, J.; Kamber, M.: Data Mining Concepts and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, USA, (2006).

[5] Kocamaz, K.: "Hastane Yönetim Bilgi Sistemlerinde Veri Madenciliği", Yüksek Lisans Tezi, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü, Konya. 2007.

[6] Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17(4), 694-701.

[7] Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert systems with Applications, 36(2), 3465-3469.

[8] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

[9] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.

[10] Inan, O., Uzer, M. S., & Yılmaz, N. (2013). A new hybrid feature selection method based on association rules and PCA for detection of breast cancer.International Journal of Innovative Computing, Information and Control, 9(2), 727-729.

[11] Lavanya, D., & Rani, D. K. U. (2011). Analysis of feature selection with classification: Breast cancer datasets.Indian Journal of Computer Science and Engineering (IJCSE), 2(5), 756-763.

[12] Jacob, S. G., & Ramani, R. G. (2011). Discovery of knowledge patterns in clinical data through data mining algorithms: multi-class categorization of breast tissue data. International Journal of Computer Applications (IJCA), 32(7), 46-53.

[13] Pach, F. P., & Abonyi, J. (2006). Association rule and decision tree based methods for fuzzy rule base generation. World Academy of Science, Engineering and Technology, 13, 45-50.

[14] Huebner, R. A. (2009). Diversity-based interestingness measures for association rule mining. Proceedings of ASBBS, 16(1).

[15] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart disease prediction system using associative classification and genetic algorithm. arXiv preprint arXiv:1303.5919.

[16] Chimieski, B. F., & Fagundes, R. D. R. (2013). Association and classification data mining algorithms comparison over medical datasets. Journal of health informatics, 5(2).

[17] Özçakır, F.C.: "Müşteri İşlemlerindeki Birlikteliklerin Belirlenmesinde Veri Madenciliği Uygulaması" Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul. 2006,

[18] Lekha, A., C. V. Srikrishna, and Viji Vinod. "Utility of association rule mining: A case study using Weka tool." Emerging Trends in VLSI, Embedded System, Nano Electronics and Telecommunication System (ICEVENT), 2013 International Conference on. IEEE, 2013.

[19] Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. 1.Frawley, eds, Knowledge Discovery in Databases, AAAIIMTT Press, Cambridge, MA.